Efficiency Considerations for Clustering the National Children's Study

by

Jeff Lehman, Warren Strauss, Louise Ryan, Steve Rust

1. Introduction

Large scale national studies that require in-person interaction with study subjects, such as in-home environmental or biological data collection, often employ multistage designs in which sufficient numbers of study subjects are clustered into small geographic regions. This clustering of study subjects is typically employed to localize data collection activities, thereby producing cost efficiencies in the data collection. These resulting financial savings could then be allocated to other study needs, such as collecting more detailed exposure data for each study subject and/or collecting other measures thought to be important for future investigations. Additionally, clustering may offer a higher degree of incentive for collecting certain types of information, such as expensive to collect community-level information or subject-specific information that is also community dependent. If a smaller number of communities are involved, collecting this type of potentially important information may be possible, whereas if a large number of communities are selected it may become infeasible or cost prohibitive to collect this data. Other possible benefits include the ability to assess within region, or within community, relationships (i.e., analyze relationships between exposures and outcomes for each community), a higher potential for collecting specialized measures (e.g., if communities are selected to correspond to an organization capable of collecting the specialized measure), and potentially higher rates of recruitment and retention as a result of communities of subjects feeling "ownership" in the study. Thus, there are many apparent benefits to clustering the design in a small number of communities or geographic regions.

While clustering of study subjects into geographic regions may provide efficiencies in cost, data collection, and the ability to collect specialized or difficult to assess measures, there may also be disadvantages to clustering subjects into a small number of regions. For example, a less diverse sample, in terms of both health outcomes and exposures, may be selected leading to lower precision in estimating parameters of interest and assessing important hypotheses. Additionally, assessment of relationships between health outcomes and community level characteristics may be less powerful as the number of communities decreases. In other words, sample clustering can result in a "cost" that is the result of a loss of information when compared to a simple random sample of the same size. The magnitude of this cost will depend on the relative amounts of within cluster variability and between cluster variability in both the exposures of interest and in the health outcome of interest. For example, if interest is in assessing the relationship between a cluster-specific exposure factor (i.e., an exposure factor that is the same for all individuals in the same cluster so that variability in the exposure only occurs between clusters), then allocating the sample in a small number of clusters will result in a loss of information; however, if interest is in assessing the relationship between a health outcome and an

exposure factor that varies significantly within a cluster and very little between clusters, then there will likely be a much smaller loss of information resulting from clustering in the design.

Since one of the NCS givens calls for the study to "… include clustering of samples to allow for efficient collection of exposure and outcome measures, and measurement of context (physical and social)", an important question that must be considered when designing the study is what degree of clustering should be employed. In other words, how many clusters are needed for the NCS to efficiently collect the necessary data while maintaining the ability to powerfully assess the hypotheses of interest in the NCS? The answer to this question depends on the characteristics of the exposures and the health outcomes considered in the NCS hypotheses, and therefore, as in many of the NCS design considerations, there is no single solution to this complex problem. This paper provides a detailed analysis of some of the statistical considerations related to sample clustering for the NCS so that an informed decision as to the appropriate number of clusters for the NCS can be made. In particular, since estimation of relationships between exposures and outcomes is of primary interest for the NCS, we explore the impacts of within and between cluster variability in outcomes and exposures on the precision (or statistical power) for assessing these relationships, and indicate how these effects vary as the number of clusters varies (given an overall sample size of 100,000). Additionally, we compare the ability for the NCS to investigate community-specific relationships (i.e., relationships within a specific cluster) with the ability for the NCS to investigate hypotheses and relationships that are National in scope (i.e., relationships that span across all clusters combined). These comparisons are made by calculating the impact of clustering and average size of clusters on the standard error of parameter estimates that relate health outcomes and exposures. For the analysis of relationships across the NCS cohort (for both binary and continuous outcomes), we consider exposures that vary within cluster as well as community-specific measures that are common to all members of a cluster/community; whereas, for the analysis of community-specific relationships, we will assume that to some degree, exposure varies within cluster and we concentrate on assessing relationships between continuous outcomes and exposures.

The remainder of this report is organized in the following manner. Section 2 presents the statistical considerations related to sample clustering and briefly outlines the methods used in assessing the impacts of sample clustering. Section 3 presents the results of applying these methods and indicates the tradeoffs between selecting a larger number of clusters with fewer people in each cluster versus a smaller number of clusters with a larger number of people in each cluster. Finally, Section 4 discusses the relevant conclusions that result from this investigation. Further details of the statistical computations can be found in the Appendix.

2. Statistical Methods and Considerations Related to Sample Clustering in the NCS

From a statistical standpoint, the advantages or disadvantages associated with sample clustering can be measured in terms of the loss or gain in precision (i.e., increased variance) in estimates of relationships of interest or the loss or gain in power for detecting these relationships. Much has been written in the sample survey literature when it comes to assessing the effect of design clustering, however, most of this literature is related to the relatively simple context where the goal is to assess the precision that a planned study might have to estimate a summary quantity, such as a mean, of a selected variable (either a health outcome variable or an exposure

variable). As mentioned above, in the context of the NCS, the situation is more complicated since estimation of relationships between exposures and outcomes is of primary interest. Thus, it is with regards to estimation of relationships that we must assess the impact of design clustering. In the following subsections we provide brief descriptions of the parameters and formulas that are utilized in assessing the impact of clustering the study design when estimating relationships between health outcomes and exposures (note that we are not evaluating impacts of unequal weighting). In particular, Section 2.1 provides information relevant to assessing the impact of clustering when estimating a relationship between a continuous outcome and a continuous exposure and Section 2.2 provides the corresponding discussion when estimating the relationship between a binary outcome and a binary exposure. Further details of the statistical computations and derivation of the relevant formulas are provided in the Appendix, while implications of the computations are explored in the results provided in Section 3.

2.1 Impact of Clustering for a Continuous Outcome and a Continuous Exposure

Suppose we are interested in exploring the relationship between an exposure and a continuous outcome, based on data from m clusters of n individuals (with $N=m \cdot n$). For simplicity, we consider a single exposure variable and let X_{ij} denote the exposure for individual j in cluster i. (In practice, of course, there will also be interest in including additional covariates.) Letting Y_{ij} be this individual's corresponding response/outcome, suppose that the model relating outcome to exposure is as follows:

$$Y_{ij} = \beta_0 + \alpha_i + \beta_1 * X_{ij} + \varepsilon_{ij} \quad \text{for } i=1,\ldots,m \text{ and } j=1,\ldots,n \tag{1}$$

where $\boldsymbol{\beta}=(\beta_0, \beta_1)^T$ are the parameters of the model, $\boldsymbol{\alpha}=(\alpha_1, \ldots, \alpha_m)^T$ is a vector of independent mean zero normally distributed random effects for each cluster (with standard deviation σ_b), and $\boldsymbol{\varepsilon}=(\varepsilon_{11}, \ldots, \varepsilon_{1n}, \ldots, \varepsilon_{mn})$ is a vector of independent mean zero normally distributed errors for each individual (with standard deviation σ_w). The model indicates that there is some relationship between the health outcome Y and the exposure factor X. The strength of this relationship depends on the value of β_1, the parameter of interest, relative to the distribution of X and the magnitude of the standard deviation for α_i and ε_{ij}. Additionally, inclusion of the random effect α allows for clustering in the health outcome that is not explained by its dependence on the exposure factor.

Under model (1) and assuming that X_{ij} follows a similar random effects model

$$X_{ij} = \mu_0 + \nu_i + \delta_{ij} \quad \text{for } i=1,\ldots,m \text{ and } j=1,\ldots,n \tag{2}$$

where ν_i and δ_{ij} are random variables with mean zero and standard deviations τ_b and τ_w, respectively, we can derive an approximate formula for the variance of $\hat{\beta}_1$, the maximum likelihood estimate of β_1, under a clustered design. In particular,

$$\mathrm{Var}(\hat{\beta}_1) \approx \frac{(\sigma_w^2 + \sigma_b^2)/(\tau_w^2 + \tau_b^2)}{m(n-1)\left(\frac{1-\lambda}{1-\rho}\right) + (m-1)\frac{1+n\lambda-\lambda}{1+n\rho-\rho}}, \tag{3}$$

where λ is the within cluster correlation of the exposure variable (or the portion of the variability in X that is explained by cluster-to-cluster variability), and ρ is the within cluster correlation of the health outcome after removing the effect of X (or the portion of the variability in Y given X that is explained by cluster-to-cluster variability) as described in Table 1. Note that the parameter ρ does not represent the marginal within cluster correlation in the health outcome, but represents the within cluster correlation in the health outcome given the exposure factor. The marginal within cluster correlation in the health outcome depends on this conditional clustering as well as the amount of clustering in the exposure factor and the strength of the relationship between the health outcome and the exposure factor.

Table 1. Summary of model and distributional assumptions.

Variable	Model	Distributions	Parameters Impacting Design Effects and Power[a]
Health Outcome	$Y_{ij} = \beta_0 + \alpha_i + \beta_1 * X_{ij} + \varepsilon_{ij}$	$\alpha_i \sim N(0, \sigma_b^2)$ $\varepsilon_{ij} \sim N(0, \sigma_w^2)$	$\rho = \frac{\sigma_b^2}{\sigma_w^2 + \sigma_b^2}$
Exposure	$X_{ij} = \mu_0 + \upsilon_i + \delta_{ij}$	$\upsilon_i \sim N(0, \tau_b^2)$ $\delta_{ij} \sim N(0, \tau_w^2)$	$\lambda = \frac{\tau_b^2}{\tau_w^2 + \tau_b^2}$

[a] ρ is the within cluster correlation in the health outcome given the exposure, and λ is the within cluster correlation in the exposure factor

The above formula for the variance of the estimate of β_1 can then be used to compare different designs in terms of their ability to estimate and detect relationships between exposures and health outcomes across the entire cohort. For example, one common measure when comparing different designs is calculating the ratio of the variance of a parameter estimate under one design to the corresponding variance under the other design. For the case where the comparison design is considered to be a simple random sample (i.e., a design with 1 subject in each of N clusters where $N=mn$) this ratio is typically referred to as a design effect. Since a simple random sample is infeasible for the NCS, and violates one of the NCS givens, in Section 3 we compute "relative design effects" as the ratio of the variance of a parameter estimate under the selected design to the corresponding variance of the parameter estimate under a design that selects 250 clusters of size 400 individuals.

If interest is in estimating the relationship between the exposure factor and the health outcome for a single cluster of individuals (instead of for the entire cohort), then $Y_{ij} = b_{0i} + b_{1i} * X_{ij} + e_{ij}$ for $j=1,\ldots,n$, and the variance of the estimate of $\hat{b}_{1i}$ is $\frac{\sigma_w^2}{\tau_w^2(n-1)}$. This formula can then be used to compute design effects (or relative design effects) and power for estimating relationships within a single cluster of individuals. (Note that in order to apply such a model the exposure factor of interest must vary within clusters so that the within cluster correlation of the exposure factor, λ, cannot be equal to 1.0.)

2.2 Impact of Clustering for a Binary Outcome and a Binary Exposure

Instead of studying the relationship between an exposure and a continuous outcome, suppose we are interested in exploring the relationship between a binary outcome and an exposure variable again based on data from clusters of individuals. In this more complex setting we consider the exposure to be dichotomous and let X_{ij} be the indicator of exposure for individual j in cluster i. Letting Y_{ij} be this individual's corresponding response, suppose that we are interested in fitting the following marginal logistic model:

$$\text{Logit}[\Pr(Y_{ij}=1|X_{ij})] = \text{Logit}(\mu_{ij}) = \beta_0 + \beta_1 X_{ij} \text{ for } i=1,\ldots,m \text{ and } j=1,\ldots,n\,. \qquad (4)$$

Generalized estimating equations (GEEs) provide an appropriate basis for analysis that accounts for both non-constant sampling probabilities, as well as for clustering of individuals (Diggle et al., 2002; Liang and Zeger, 1986). In Section A.2 of the Appendix we use these estimating equations to derive a formula for the variance of the estimate of β_1 under the clustered design. Due to the mean-variance relationship for binomial data, derivation of this formula is more complex and depends on a number of factors, including:

- m = the number of clusters,
- n = the number of individuals in each cluster,
- ρ = the within cluster correlation in the health outcome,
- λ = the within cluster correlation in the exposure factor,
- p_1 = the probability of exposure,
- μ_0 = the probability of disease for unexposed individuals, and
- β_1 = the log-odds ratio describing the strength of the relationship between outcome and exposure.

We refer the reader to the Appendix for further details on calculating the variance of the estimate of β_1 (the log-odds ratio) under a clustered design.

3. Results

In this Section, we display and discuss a series of figures representing the impacts of clustering for estimating relationships between continuous outcomes and continuous exposures across the entire cohort and within each cluster (Section 3.1). Additionally, figures displaying the impacts of estimating relationships between binary outcomes and binary exposures across the entire cohort are also provided (Section 3.2). In general, these impacts are evaluated in terms of design effects for estimating the relationship of interest and/or in terms of the power to detect a relationship of a specified size. Instead of providing a design effect that is the ratio of the variance of the parameter estimate under the selected design to that of a simple random sample, we instead compute a relative design effect that is the ratio of the variance of the parameter estimate under the selected design to the corresponding variance under a design with 250 clusters of size 400 (i.e., since a simple random sample is infeasible in this setting).

3.1 Results for Continuous Outcomes and Continuous Exposure Factors

First concentrating on the impacts of clustering in estimating the relationship between a continuous outcome and a continuous exposure across the entire cohort, Figures 1 and 2 display the relative design effect (ratio of parameter estimate variance under the selected design to a design with 250 clusters) for estimation of the relationship and the power for detecting the relationship, respectively. In particular, since the relative design effect for estimation of relationships across the entire cohort depends on the within cluster correlation in exposure (λ) and on the within cluster correlation in the health outcome after removing the effect of exposure (ρ), Figure 1 displays the relative design effect as a function of λ, for different values of ρ and for different numbers of clusters (assuming an overall sample size of 100,000 individuals and a reference design with 250 clusters of size 400). The figure indicates that as ρ increases the impact of clustering in the exposure factor (λ) changes, with the case of a cluster-specific exposure factor ($\lambda=1$) having the largest loss of information as a result of design clustering. However, note that in general, even for exposure factors with a large degree of within cluster correlation (e.g., λ less than 0.5), the relative design effect is very close to one in all cases.

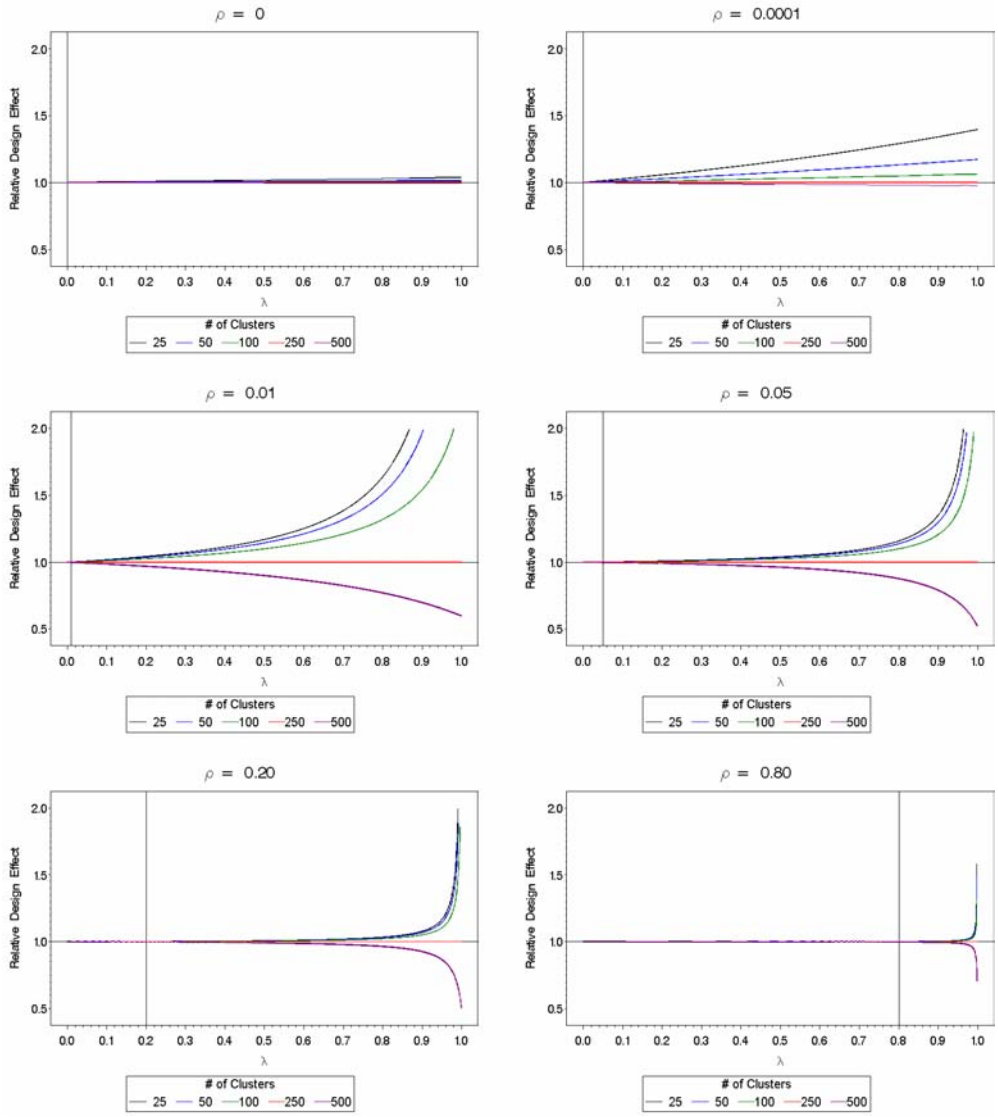

Figure 1. Relative design effects (with respect to a reference design with 250 clusters of size 400) for estimating relationships across the entire cohort.

To translate these relative design effects for assessing an exposure/outcome relationship across the entire cohort to estimates of the power to detect that relationship, the magnitude or strength of the relationship must be specified. Since very weak relationships will be undetectable with any design (i.e., the power will be equal to the significance level of the test regardless of the design) and very strong relationships are detectable with any design (i.e., the power will be equal to 1.0 regardless of the design), Figure 2 displays the power to detect a relationship that has a magnitude which is detectable with 80% power under a design with 250 clusters of 400 individuals (note that this magnitude may change as a function of ρ and λ). By so doing, the figure indicates the loss of (or gain in) power when attempting to detect a relationship that would be detectable with sufficient power if a 250 cluster design were adopted (i.e., treating the 250 cluster design as a reference design). As seen in the relative design effects, there is very little loss of power as a result of design clustering when the within cluster correlation in the

health outcome is very small (e.g., ρ less than 0.0001) and/or the within cluster correlation in the exposure factor is small (e.g., λ less than 0.5). On the other hand, for assessing relationships between health outcomes with large within cluster correlation (e.g., ρ larger than 0.01) and cluster-specific exposure factors (λ=1), the loss in power when going from a 250 cluster design to a 25 cluster design can be on the order of 60% (i.e., from 80% power to approximately 20% power).

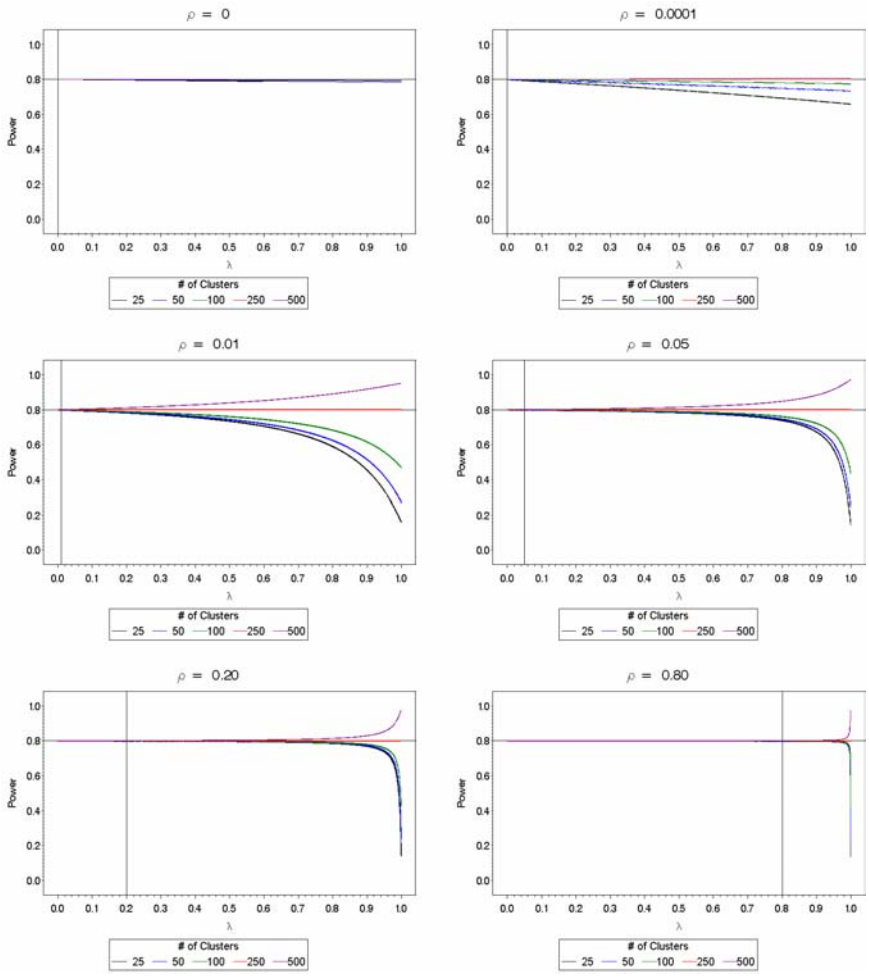

Figure 2. Comparison of power to detect an exposure/outcome relationship across the entire cohort.

Assuming that to some degree exposure varies within cluster (so that λ is strictly less than 1.0), estimation of a community-specific relationship is feasible. As indicated in Section 2.1, the relative design effect for estimating a relationship within a cluster will depend only on the number of clusters, or the number of individuals within each cluster. Thus, Table 2 displays the relative design effect, with respect to a reference design with 250 clusters of size 400, as a function of the number of clusters, and assuming an overall sample size of 100,000 individuals. As is intuitively reasonable, the table demonstrates than when the goal is estimation of relationships within a single cluster, a larger number of individuals in each cluster will lead to a more optimal design (i.e., smaller relative design effects). Translating these cluster-specific relative design effects into estimates of the power to detect a relationship within a cluster, Table 2 also displays the power to detect a relationship with a magnitude that is detectable with 80%

power under a design with 250 clusters of 400 individuals (i.e., a relationship with a magnitude such that the effect divided by the standard error of the effect under this design is approximately 2.8). As in the relative design effects, larger power to detect a within cluster relationship is realized when a larger number of individuals are sampled in each cluster.

Table 2. Comparison of relative design effects (with respect to a reference design with 250 clusters of size 400) and power when estimating an exposure/outcome relationship within a single cluster.

Number of Clusters	Individuals Per Cluster	Relative Design Effect for Estimating Within Cluster Relationships	Power for Estimating Within Cluster Relationships
25	4000	0.100	1.000
50	2000	0.200	1.000
100	1000	0.399	0.993
250	400	1.000	0.800
500	200	2.005	0.507

3.2 Results for Binary Outcomes and Binary Exposure Factors

Displaying design effects, or statistical power, when estimating relationships between a binary exposure factor and a binary health outcome is slightly more complicated since the variance of the estimate that describes the relationship between the disease and the exposure (i.e., the log-odds ratio) depends on a larger number of factors. In particular, as described in Section 2.2, uncertainty in the relationship (as estimated by the variance of the log-odds ratio) depends on the number of clusters (m), the number of individuals in each cluster (n), the within cluster correlation in the health outcome (ρ), the within cluster correlation in exposure (λ), the probability of exposure (p_1), the probability of the disease for unexposed individuals (μ_0), and the strength of the relationship between outcome and exposure denoted by the odds ratio [$OR=\exp(\beta_1)$]. Since displaying figures that allow all of these factors to vary would result in a large number of illustrations, we focus on just a few settings of the above factors that are relevant to the NCS. In terms of disease prevalence, we focus on two examples, one representing a relatively rare outcome, such as autism or schizophrenia, with $\mu_0=0.005$ (i.e., a 0.5% chance of disease for unexposed individuals), and the other representing a more common outcome, such as asthma or obesity, with $\mu_0=0.05$ (i.e., a 5% chance of disease for unexposed individuals). For both of these cases we will assume that the probability of exposure is 0.10 so that approximately 10% of the population is exposed, and we allow all other factors (m, n, λ, ρ, and the OR) to range over a reasonable set of values. In particular, as in Section 3.1, we evaluate designs with 25, 50, 100, 250, and 500 clusters of individuals (always assuming a total sample size of 100,000), and allow ρ to take values of 0.001, 0.01, and 0.10 representing a reasonable range of possible within cluster correlations in subject-specific binary health outcomes and λ to take values of 0.01, 0.1, and 1.0 (with $\lambda=1.0$ representing a cluster-specific exposure variable).

Figure 3 displays the power to detect the exposure/outcome relationship as a function of the odds ratio (the strength of the exposure/outcome relationship) for a relatively common health outcome that has a response probability of 5% for unexposed individuals. As in Section 3.1, there is very little loss of power when comparing designs with a smaller number of clusters if the

within cluster correlation in the health outcome and the within cluster correlation in exposure are small (e.g., λ=0.01 and ρ=0.001 or 0.01); whereas, when ρ and λ increase selecting fewer clusters results in a loss of power with the magnitude of this loss depending on the specified odds ratio. Table 3 summarizes the information in Figure 3 by displaying the odds ratios that are detectable with at least 80% power (increments of 0.1) under the different designs and different degrees of clustering in exposure and health outcome.

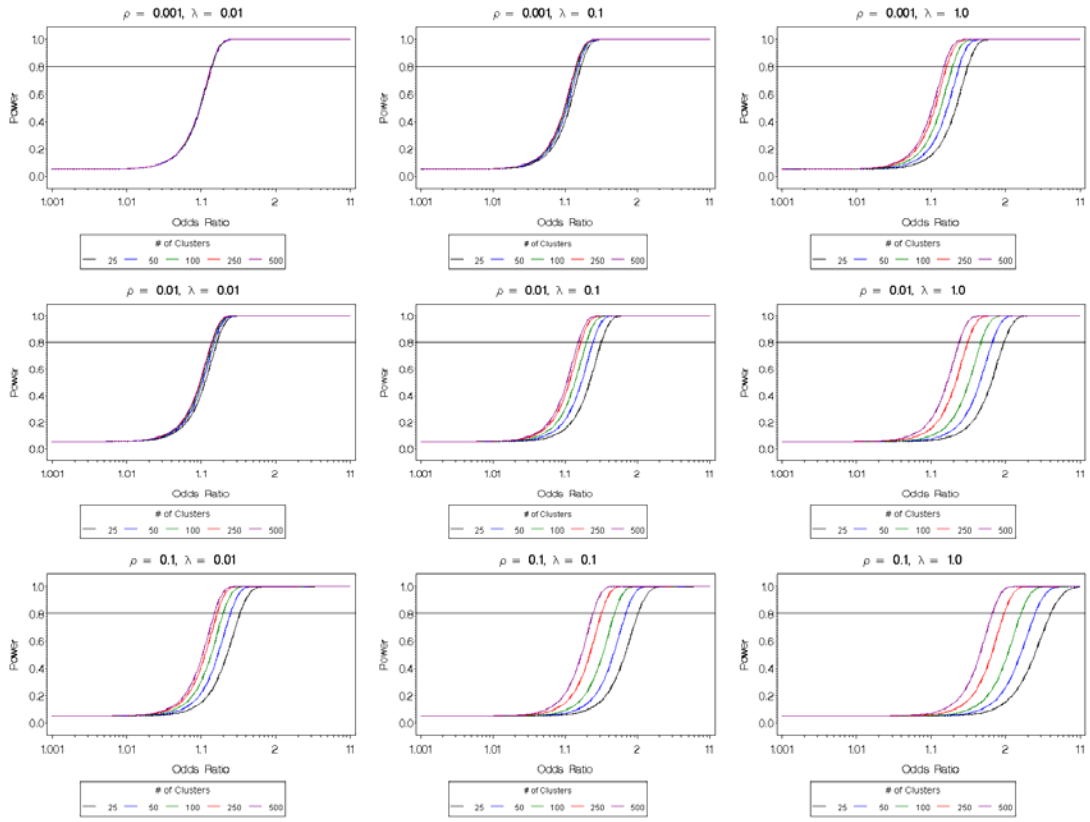

Figure 3. Comparison of power to detect a binary exposure/binary outcome relationship across the entire cohort for a relatively common health outcome (probability of disease for unexposed individuals of 0.05) and an exposure factor that occurs in 10% of the population.

Table 3. Odds ratios detectable with 80% power for a common health outcome (probability of disease for unexposed individuals of 0.05) and an exposure factor that occurs in 10% of the population.

ρ	λ	Odds Ratio Detectable with 80% Power				
		25 Clusters	50 Clusters	100 Clusters	250 Clusters	500 Clusters
0.001	0	1.2	1.2	1.2	1.2	1.2
	0.01	1.2	1.2	1.2	1.2	1.2
	0.1	1.2	1.2	1.2	1.2	1.2
	1	1.4	1.3	1.2	1.2	1.2
0.01	0	1.2	1.2	1.2	1.2	1.2
	0.01	1.2	1.2	1.2	1.2	1.2
	0.1	1.4	1.3	1.2	1.2	1.2
	1	2.0	1.7	1.5	1.4	1.3
0.1	0	1.2	1.2	1.2	1.2	1.2
	0.01	1.4	1.3	1.2	1.2	1.2
	0.1	2.1	1.7	1.5	1.4	1.3
	1	5.1	3.5	2.7	2.0	1.7
0.5	0	1.3	1.2	1.2	1.2	1.2
	0.01	2.0	1.6	1.4	1.3	1.2
	0.1	4.2	2.9	2.2	1.7	1.5
	1	>10	>10	5.9	3.5	2.7

Figure 4, on the other hand, displays the power to detect the exposure/outcome relationship as a function of the odds ratio (the strength of the exposure/outcome relationship) for a rare health outcome that has a response probability of 0.5% for unexposed individuals, and Table 4 displays the corresponding odds ratios that are detectable with at least 80% power. Comparing these plots to the plots displayed in Figure 3, larger odds ratios are necessary in order to detect a relationship due to the smaller probability of disease; however, in terms of evaluating the impact of clustering on the power to detect relationships similar conclusions are apparent. In particular, the impact of clustering on statistical power and/or the odds ratios detectable with 80% power are very small when λ and/or ρ are small, and increase as these values get larger.

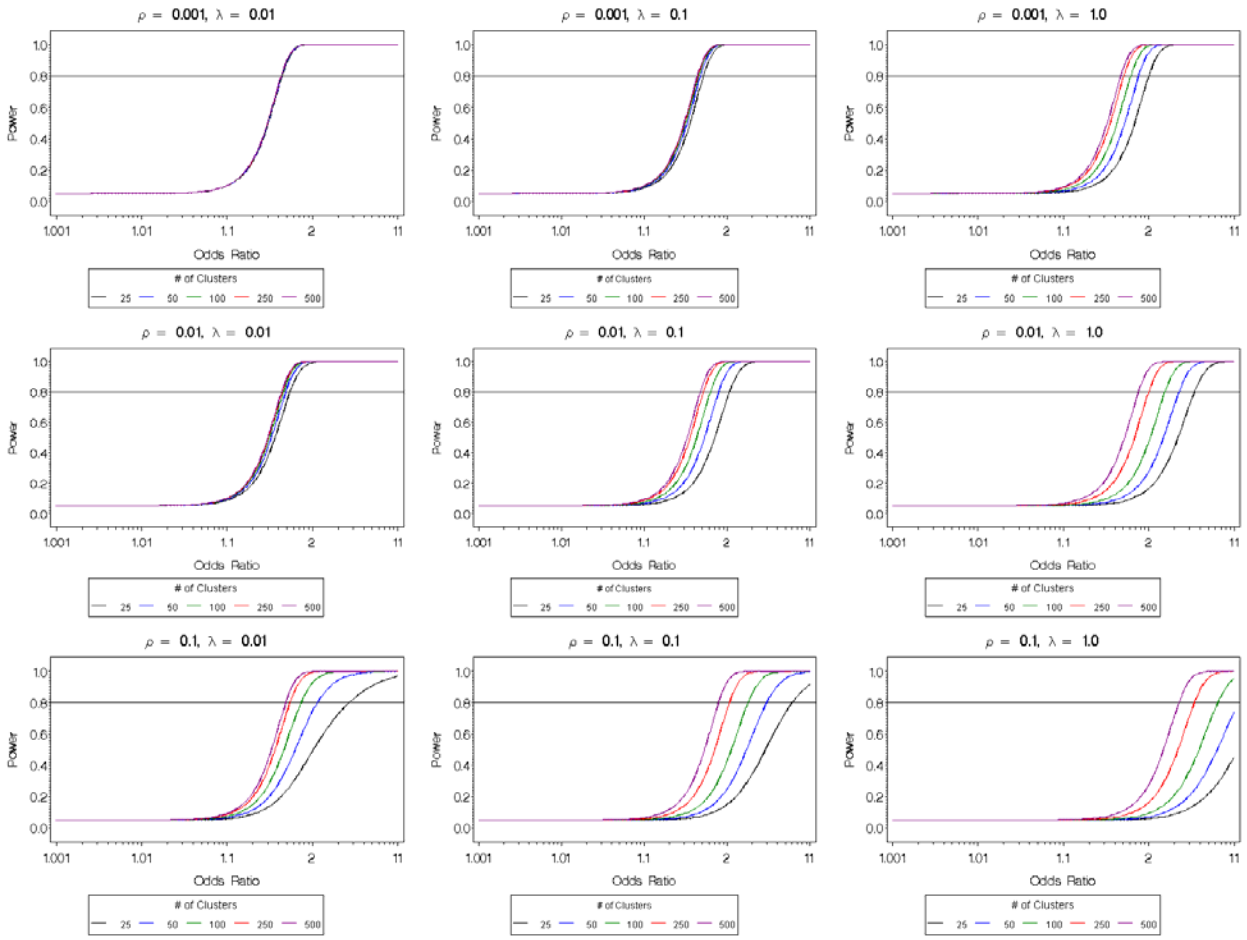

Figure 4. **Comparison of power to detect a binary exposure/binary outcome relationship across the entire cohort for a relatively common health outcome (probability of disease for unexposed individuals of 0.005) and an exposure factor that occurs in 10% of the population.**

Table 4. Odds ratios detectable with 80% power for a common health outcome (probability of disease for unexposed individuals of 0.005) and an exposure factor that occurs in 10% of the population.

ρ	λ	Odds Ratio Detectable with 80% Power				
		25 Clusters	50 Clusters	100 Clusters	250 Clusters	500 Clusters
0.001	0	1.5	1.5	1.5	1.5	1.5
	0.01	1.5	1.5	1.5	1.5	1.5
	0.1	1.6	1.5	1.5	1.5	1.5
	1	2.1	1.8	1.7	1.6	1.5
0.01	0	1.5	1.5	1.5	1.5	1.5
	0.01	1.6	1.5	1.5	1.5	1.5
	0.1	2.1	1.8	1.7	1.6	1.5
	1	4.5	3.3	2.6	2.1	1.8
0.1	0	2.9	1.7	1.6	1.5	1.5
	0.01	3.7	2.2	1.8	1.6	1.5
	0.1	7.1	4.0	2.8	2.1	1.8
	1	>10	>10	7.5	4.5	3.3
0.5	0	>10	>10	4.3	1.7	1.6
	0.01	>10	>10	5.1	2.2	1.8
	0.1	>10	>10	9.2	4.0	2.8
	1	>10	>10	>10	>10	7.5

In terms of evaluating the impact of clustering when estimating a community-specific relationship between a binary outcome and a binary exposure (assuming that to some degree exposure varies within cluster), we envision the results will be the same as those displayed in Table 2.

4. Conclusions

The results in Section 3 demonstrate the statistical considerations relevant to clustering the NCS design in a number of geographic communities, and provide comparisons of the loss or gain in statistical efficiency (in terms of relative design effects or statistical power) when estimating relationships between exposures and health outcomes. In particular, designs with 25, 50, 100, 250, and 500 clusters (each assuming a total sample size of 100,000 individuals) were compared for hypotheses involving estimation of relationships between outcomes (binary and continuous) and exposures (binary and continuous) across the entire cohort, and for hypotheses involving estimation of relationships between continuous outcomes and continuous exposures within a single cluster (i.e., cluster-specific relationships).

In terms of statistical efficiency when estimating relationships across the entire cohort, the loss or gain associated with a clustered design depends on the characteristics of the exposures and the health outcomes of interest. In particular, the amount of within cluster correlation in the health outcome (denoted as ρ in the results of Section 3) along with the amount of within cluster

correlation in the exposure factor (denoted as λ in the results of Section 3) has significant influence on the loss/gain in statistical efficiency resulting from clustering the design. In general, when these correlations are small, so that the health outcome and the exposure factor primarily varies within cluster, there are little to no differences between designs with 500 clusters of 200 individuals and designs with 25 clusters of 4000 individuals (i.e., there is little loss of information resulting from clustering the design). On the other hand, for cluster-specific exposure factors (i.e., an exposure factor that only varies between clusters and is constant within clusters) and for larger values of the within cluster correlation in the health outcome the impacts of design clustering are less trivial.

Therefore, in order to further quantify the impacts of clustering and to indicate the tradeoffs inherent in selecting a design with a larger or smaller number of clusters, we must consider reasonable values for the within and between cluster variability in exposures and health outcomes of interest in the NCS. The NCS primary health outcomes (asthma, obesity, pregnancy outcomes, neurodevelopmental and behavioral outcomes, injury outcomes), are evaluated on a subject-specific level and will arguably have a relatively small degree of geographic clustering (i.e., small amount of within cluster correlation in health outcomes). For exposure factors, on the other hand, there are certainly exposure factors that vary primarily within cluster (e.g., personal activity levels, exposure to pesticides, exposure to mediators of inflammation, etc.) as well as factors that vary primarily, or only, between clusters (e.g., community-level $PM_{2.5}$ concentrations, community-level housing variables, etc.); however, since most of the NCS core hypotheses call for subject-specific analyses, and since many of the variables that vary only between clusters may be interacted with subject-specific variables to form an "exposure" variable (e.g., interaction of community-level $PM_{2.5}$ concentrations and subject-specific activity patterns to form a $PM_{2.5}$ exposure metric), it may be reasonable to assume that most of the primary exposure covariates will have significant within-cluster variability.

Thus, under the assumption that there is a relatively small degree of within cluster correlation in the health outcomes and exposure factors of primary interest in the NCS (i.e., a significant portion of the variability in the health outcome and the exposure occurs within a cluster), there appears to be little loss of statistical efficiency in estimating relationships across the entire cohort (small relative design effects and little loss of power) when comparing designs with 50 or 100 clusters to designs with 250 or 500 clusters. Balanced against other considerations related to sample clustering, such as the financial efficiencies of data collection in a smaller number of geographic regions, this suggests little need for selecting a large number of clusters in the NCS. This is not meant to suggest that there is no advantage to selecting a design with a larger number of clusters, but rather we are suggesting that there does not appear to be a *significant* statistical advantage to selecting a design with 250 or 500 clusters as compared to a design with 50 or 100 clusters. This makes the overall advantage (i.e., when considering all factors, not just statistical, related to sample clustering) to a larger number of clusters less attractive. Adding further argument for a smaller number of clusters is the greater statistical efficiency for estimating within-cluster relationships given a design with fewer clusters (and a large number of individuals in each cluster).

Of course, it must be noted that the above conclusions rely on the assumption that there is a small degree of within cluster correlation in the health outcomes and the exposure factors that

are of primary interest in the NCS. That is not to suggest that there are no exposure factors or health outcomes that primarily vary between clusters, or that there is no interest in evaluating these types of outcomes or exposures. Certainly there are community-level effects, such as average housing age or median household income, that are of interest to a large portion of the scientific community. For evaluating only these types of exposure factors or evaluating cluster-specific health outcomes, there are clear advantages to selecting a larger number of clusters. Additionally, given that the NCS is called to serve as a resource for future hypotheses and studies, there are definite arguments for a design that selects a larger number of clusters.

As suggested in Section 1, in addition to the statistical advantages/disadvantages relevant to determining the appropriate number of clusters for the NCS, there are a variety of other important factors (e.g., cost considerations, data availability considerations, etc.) influencing this decision. While the statistical efficiency results discussed in this report did not incorporate these other factors (for example we did not consider the possibility that a 25 cluster design may provide a larger amount of data or a higher degree of study subject retention due to financial savings of data collection being rerouted to other Study activities), the ultimate decision of the number of clusters must consider them in conjunction with the statistical considerations related to sample clustering. The results presented above provide a detailed analysis (and a means of providing further evaluation and assessment of power for a selected hypothesis under any clustered design) of the statistical impacts associated with sample clustering in the NCS. By combining these considerations with the other factors influencing the determination of the number of clusters, an informed decision as to the appropriate number of clusters for the NCS can be made.

5. References

Diggle, P., Heagerty, P., Liang, K-Y., Zeger, S. (2002). Analysis of Longitudinal Data. Oxford University Press.

Liang, K.Y. and Zeger S.L. (1986). Longitudinal data analysis using generalized linear models. Biometrika. 73:13-22.

Appendix A - Statistical Methods: Efficiency Considerations for Clustering the National Children's Study

A.1 Continuous Outcome and Continuous Exposure Factor

To compute the precision when estimating a relationship between an exposure and an outcome in a clustered design, we begin from first principles. Letting X_{ij} be the exposure for individual j in cluster i, and Y_{ij} this individual's corresponding response/outcome, suppose that the relationship between the outcome and exposures is as follows:

$$Y_{ij} = \beta_0 + \alpha_i + \beta_1 * X_{ij} + \varepsilon_{ij} \quad \text{for } i=1,\ldots,m \text{ and } j=1,\ldots,n$$

or in matrix notation

$$\mathbf{Y} = \mathbf{F}\beta + \mathbf{Z}\alpha + \varepsilon$$

where $\boldsymbol{\beta}=(\beta_0, \beta_1)^T$ are the parameters of the model, $\boldsymbol{\alpha}=(\alpha_1, \ldots, \alpha_m)^T$ is a vector of independent mean zero normally distributed random effects for each cluster (with standard deviation σ_b), $\boldsymbol{\varepsilon}=(\varepsilon_{11}, \ldots, \varepsilon_{1n}, \ldots, \varepsilon_{mn})$ is a vector of independent mean zero normally distributed errors for each individual (with standard deviation σ_w), and **F** and **Z** are the fixed and random effects design matrices, respectively. The model indicates that there is some relationship between the health outcome Y and the exposure factor X. The strength of this relationship depends on the value of β_1, the parameter of interest, relative to the distribution of X and the magnitude of the standard deviation for α_i and ε_{ij}. Additionally, inclusion of the random effect α allows for clustering in the health outcome that is not explained by its dependence on the exposure factor.

Given the values of the exposure variable (i.e., given the X_{ij}'s), the maximum likelihood estimate for $\boldsymbol{\beta}$ and it's corresponding variance is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{F}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{F})^{-1}\mathbf{F}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{Y} \qquad \text{and} \qquad \text{Var}(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) = (\mathbf{F}^{\mathbf{T}}\mathbf{V}^{-1}\mathbf{F})^{-1},$$

where **V** represents the variance-covariance matrix of the vector **Y**. Focusing on the estimate of β_1, the parameter describing the relationship between exposure and outcome, and simplifying the above expression we have

$$\text{Var}(\hat{\beta}_1 \mid \mathbf{X}) = \frac{(n\sigma_b^2 + \sigma_w^2)\sigma_w^2}{n\sigma_b^2 \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i\bullet})^2 + n\sigma_w^2 \sum_{i=1}^{m}(\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})^2},$$

where $\bar{x}_{i\bullet}$ represents the mean of $(x_{i1}, \ldots, x_{in})$, and $\bar{x}_{\bullet\bullet}$ represents the mean exposure over all individuals. The formula can be derived by noting that **F** is an nm by 2 matrix with one's in the

first column and the x_{ij}'s in the second column, and $\mathbf{V}$ is a block diagonal matrix with m blocks of size n by n each having $\sigma_b^2 + \sigma_w^2$ on the diagonal and σ_b^2 on the off diagonals. In other words, $\mathbf{V} = \mathbf{V_1} \otimes \mathbf{I_m}$ where $\mathbf{V_1} = \sigma_w^2 \mathbf{I_n} + \sigma_b^2 \mathbf{J_n}$, $\mathbf{I_p}$ is the p by p identity matrix and $\mathbf{J_p}$ is a p by p matrix of one's. In this case, $\mathbf{V}^{-1} = \mathbf{V}_1^{-1} \otimes \mathbf{I_m}$ and $\mathbf{V}_1^{-1} = \sigma_w^{-2}\mathbf{I_n} - \frac{\sigma_b^2}{\sigma_w^2(n\sigma_b^2 + \sigma_w^2)}\mathbf{J_n}$, which can be used to derive the above formula.

This provides a formula for the variance of the estimate of the relationship between the health outcome variable and the exposure variable. To compute the marginal variance of $\hat{\beta}_1$ we can remove the dependence on the exposure variable, $\mathbf{X}$, by specifying a distribution for $\mathbf{X}$ and taking expectations over this distribution. For example, assuming that

$$X_{ij} = \mu_0 + \nu_i + \delta_{ij} \quad \text{for } i=1,\ldots,m \text{ and } j=1,\ldots,n$$

where ν_i and δ_{ij} have mean zero and standard deviations τ_b and τ_w, respectively. Using a taylor series expansion we have

$$\mathrm{E}_{\mathrm{X}}[\mathrm{Var}(\hat{\beta}_1 \mid \mathbf{X})] \approx \frac{1}{m(n-1)\frac{\tau_w^2}{\sigma_w^2} + (m-1)\frac{n\tau_b^2 + \tau_w^2}{n\sigma_b^2 + \sigma_w^2}} = \frac{(\sigma_w^2 + \sigma_b^2)/(\tau_w^2 + \tau_b^2)}{m(n-1)\left(\frac{1-\lambda}{1-\rho}\right) + (m-1)\frac{1+n\lambda-\lambda}{1+n\rho-\rho)}},$$

where $\lambda = \frac{\tau_b^2}{\tau_w^2 + \tau_b^2}$ is the within cluster correlation of the exposure variable and $\rho = \frac{\sigma_b^2}{\sigma_w^2 + \sigma_b^2}$ is the within cluster correlation of the health outcome after removing the effect of X. Note that the parameter ρ does not represent the marginal within cluster correlation in the health outcome, but represents the within cluster correlation in the health outcome given the exposure factor. The marginal within cluster correlation in the health outcome depends on this conditional clustering as well as the amount of clustering in the exposure factor and the strength of the relationship between the health outcome and the exposure factor.

One common measure when comparing different designs is in the design effect which represents the ratio of the parameter estimate variance under the selected design to the parameter estimate variance under a single common design (e.g., a simple random sample design). Based on the above formula, the design effect resulting from clustering the subjects in m clusters of size n when estimating a relationship across the entire cohort (i.e., the ratio of the variance of $\hat{\beta}_1$ for a design with 1 subject for each of N clusters to a design with n subjects for each of m clusters such that $N=mn$) can be written as

$$\mathrm{DE}_{\mathrm{C}} \approx \frac{(mn-1)}{m(n-1)\left(\frac{1-\lambda}{1-\rho}\right) + (m-1)\frac{1+n\lambda-\lambda}{1+n\rho-\rho)}}.$$

On the other hand, if interest were in estimating the relationship between the exposure factor and the health outcome for a single cluster of individuals, or each cluster separately, (i.e., $Y_{ij} = b_{0i} + b_{1i}*X_{ij} + e_{ij}$ for $j=1,...,n$) then the variance of the estimate of $\hat{b}_{1i}$ would be $\frac{\sigma_w^2}{\tau_w^2(n-1)}$, and the ratio of this variance for a design with m clusters of size n to a design with 1 cluster of size mn is then $\frac{mn-1}{n-1}$, representing the impact of clustering when estimating relationships within a cluster. (Note that in order to apply such a model the exposure factor of interest must vary within clusters so that the within cluster correlation of the exposure factor, λ, cannot be equal to 1.0.)

Figures A.1 and A.2 display design effects for estimation of relationships across the entire cohort and estimation of relationships within a single cluster, respectively. In particular, since the design effect for estimation of relationships across the entire cohort depends on the within cluster correlation in exposure, λ, and on the within cluster correlation in the health outcome after removing the effect of exposure, ρ, Figure A.1 displays the design effect as a function of λ, for different values of ρ, and for different numbers of clusters (assuming an overall sample size of 100,000 individuals). On the other hand, since the design effect for estimating a relationship within a cluster depends only on the number of clusters (or the number of individuals within each cluster), Figure A.2 simply displays the design effect as a function of the number of clusters (again, assuming an overall sample size of 100,000 individuals).

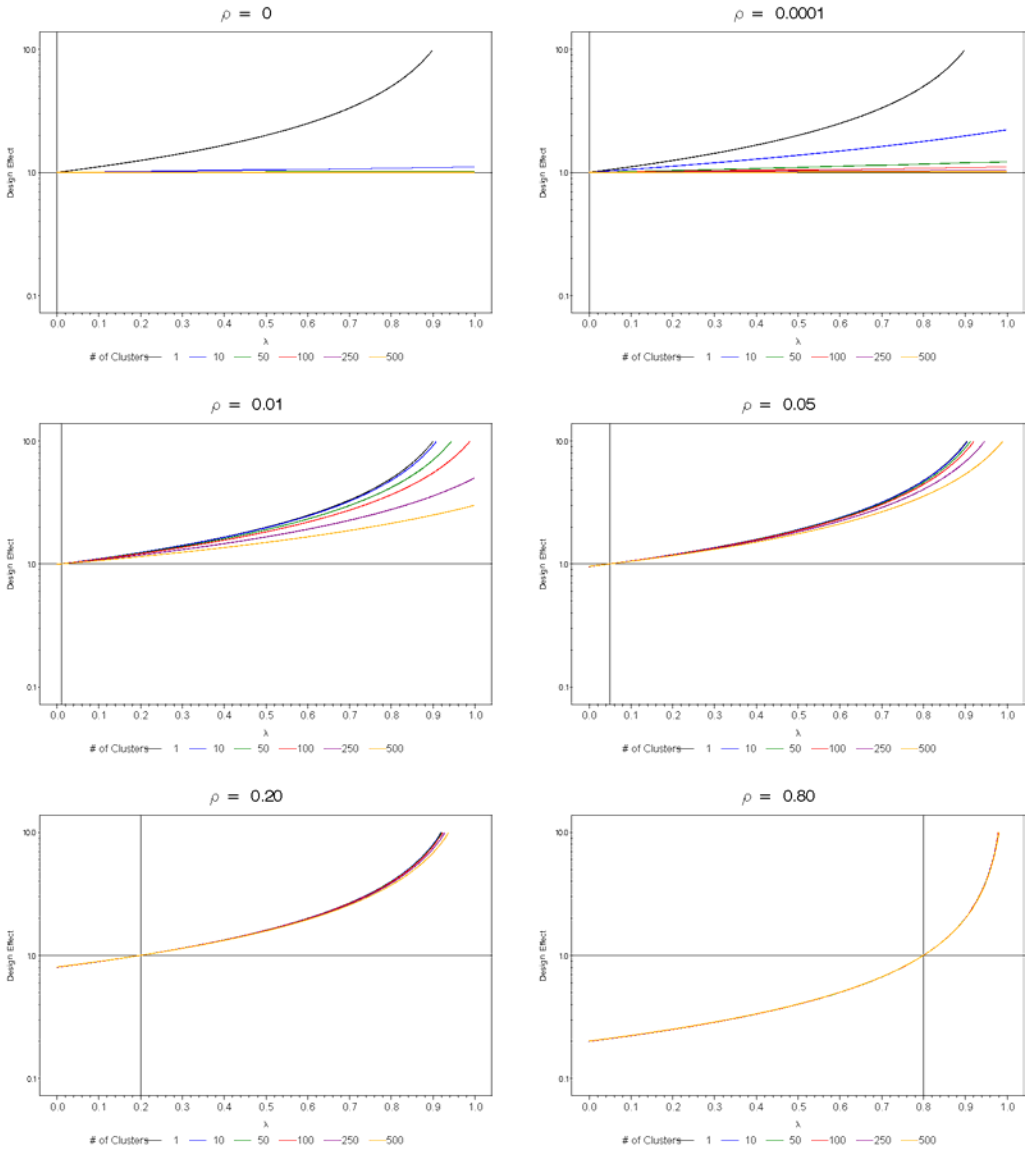

Figure A.1. Design effects (with respect to a simple random sample) for estimating relationships across the entire cohort.

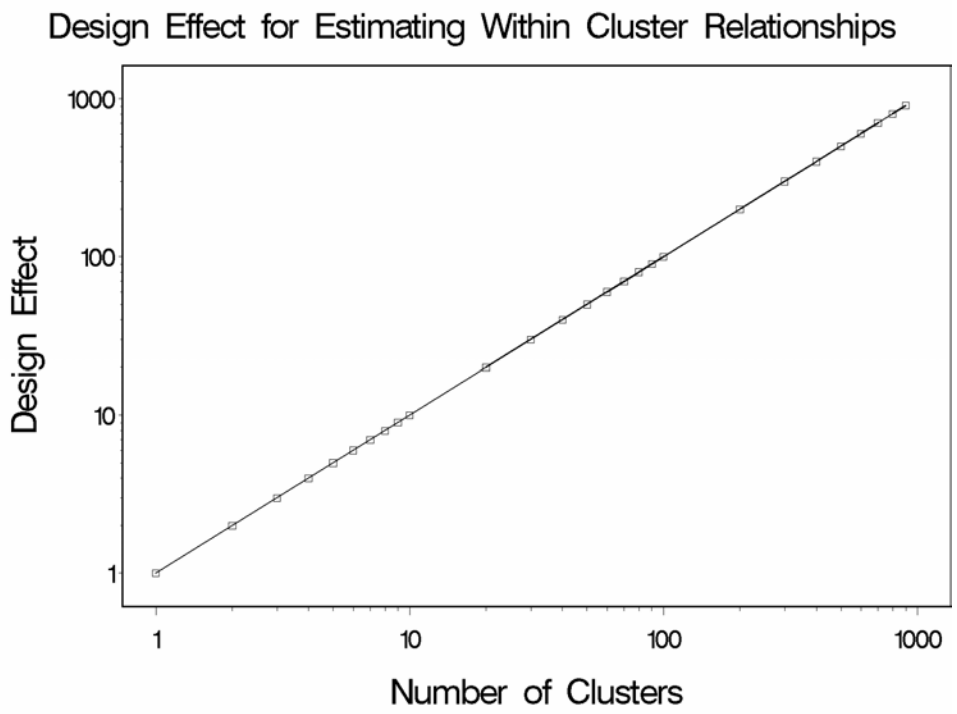

Figure A.2. Ratio of variance for estimating a within cluster relationship for a design with the specified number of clusters to the corresponding variance for a design with a single cluster.

To translate the design effects for assessing an exposure/outcome relationship across the entire cohort to estimates of the power to detect the relationship, Figure A.3 displays a series of power curves for assessing a relationship that has a magnitude which would be detectable with 80% power under a design with no clustering (i.e., a relationship with a magnitude such that the effect divided by the standard error of the effect under a design with no clustering is approximately 2.8).

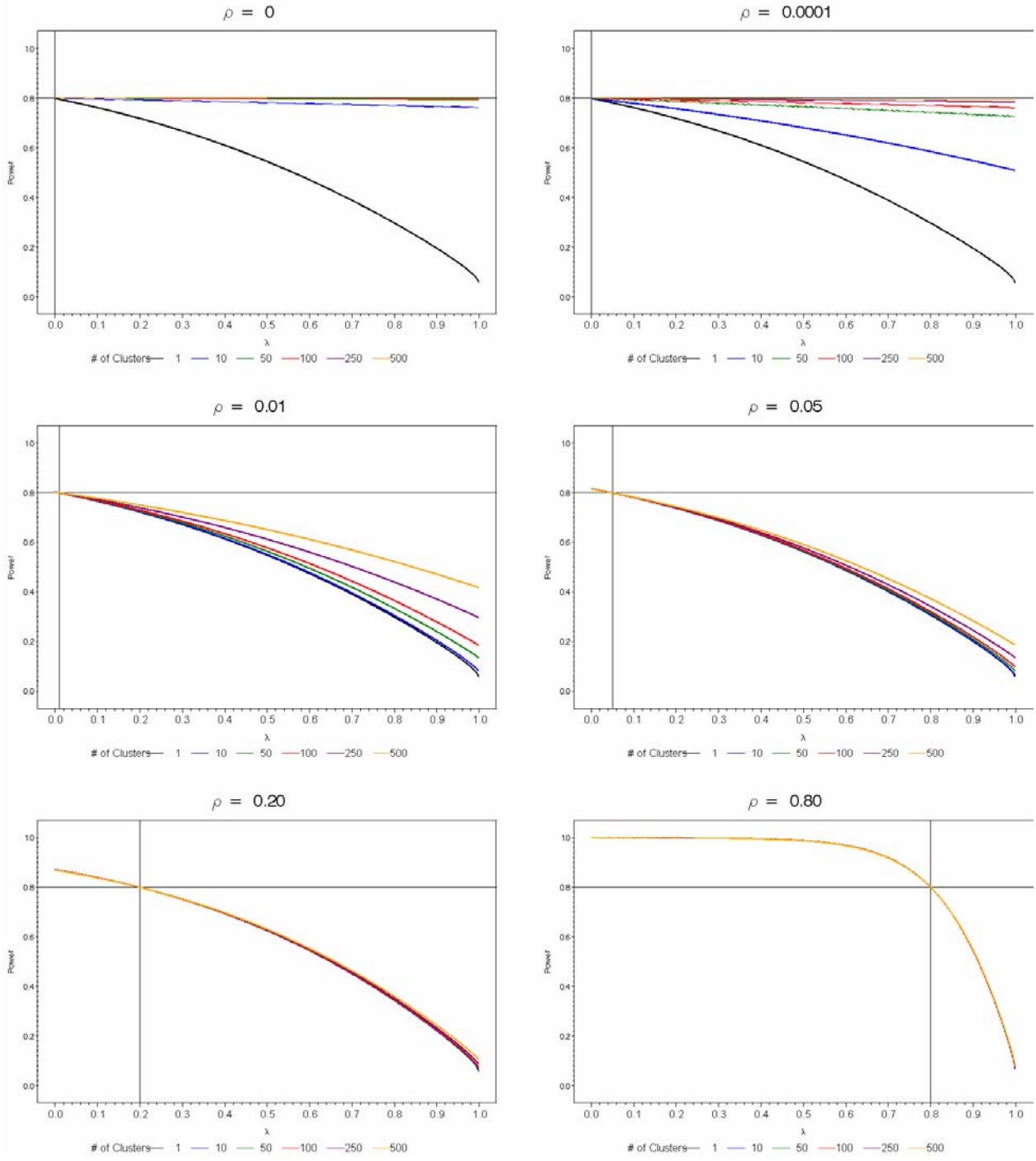

Figure A.3. Comparison of power to detect an exposure/outcome relationship across the entire cohort.

Finally, to translate the design effects for assessing an exposure/outcome relationship within a single cluster to estimates of the power to detect the relationship, Figure A.4 displays

the power curve for assessing a relationship that has a magnitude which would be detectable with 80% power under a design that selects all individuals in a single cluster (i.e., a relationship with a magnitude such that the effect divided by the standard error of the effect under a design with all individuals in one cluster is approximately 2.8).

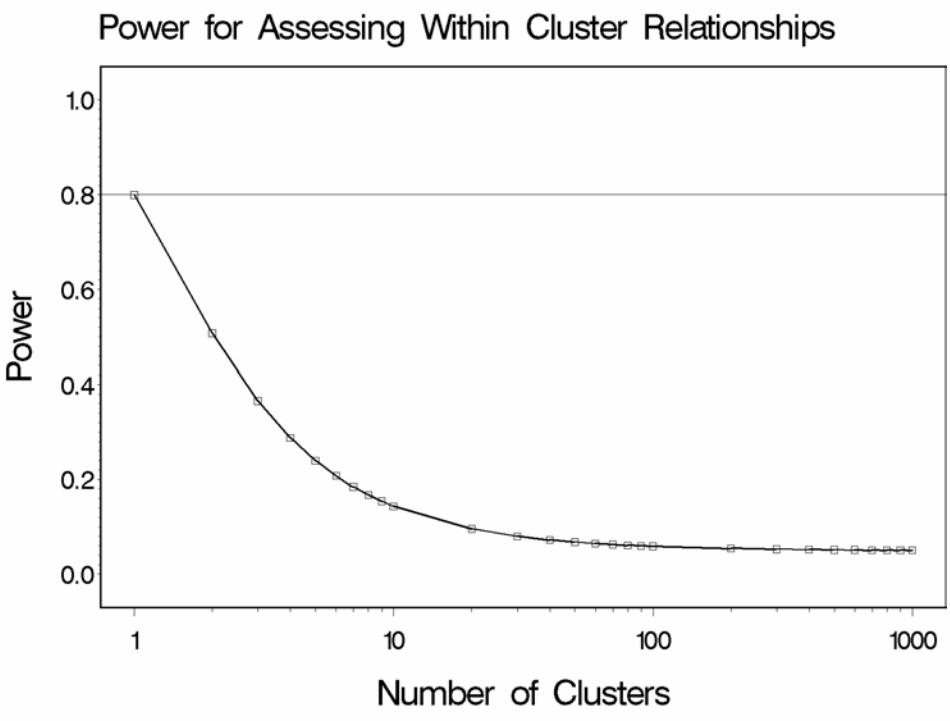

Figure A.4. Comparison of power to detect an exposure/outcome relationship within a single cluster.

A.2 Binary Outcome and Binary Exposure Factor

As above, let X_{ij} be the exposure for individual j in cluster i and let Y_{ij} be this individual's corresponding response, and suppose also that we are interested in fitting the following marginal logistic model:

$$\text{Logit}[\Pr(Y_{ij}=1|X_{ij})] = \text{Logit}(\mu_{ij}) = \beta_0 + \beta_1 X_{ij.} \quad (1)$$

(In practice, of course, there will also be interest in including additional covariates and risk factors.) Generalized estimating equations (GEEs) provide an appropriate basis for analysis that accounts for both non-constant sampling probabilities, as well as for clustering of individuals (Diggle et al., 2002; Liang and Zeger, 1986). Letting w_{ij} be the sampling weight for the j^{th} individual in the i^{th} cluster (generally, this will be the inverse of their selection probability), a suitable estimating equation for the unknown parameter $\beta=(\beta_0,\beta_1)^T$ is

$$U(\beta) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij} \begin{pmatrix} 1 \\ X_{ij} \end{pmatrix} (Y_{ij} - \mu_{ij}) = 0 \,, \tag{2}$$

where m is the number of clusters, n_i is the number of subjects in cluster i, and μ_{ij} is the mean response for individual j in the i^{th} cluster. (Note that interest here in this paper is in assessing the impact of clustering, not the impact of unequal weighting. Thus, in the results presented we assume all w_{ij}'s are equal; however, in these calculations we provide the more generic derivation of these results by incorporating the w_{ij}'s.) Standard estimating equations theory can be used to establish that the variance of the parameter estimates, $\hat{\beta}$, is

$$\text{Var}(\hat{\beta}) = B^{-1} A (B^T)^{-1} \tag{3}$$

where B is the matrix of partial derivatives of $U(\beta)$ and A is the variance of $U(\beta)$. This is the calculation automatically performed in software such as SUDAAN or SAS PROC GENMOD (if the empirical variance option is invoked). It is relatively straightforward to show that

$$B = \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij} \mu_{ij} (1 - \mu_{ij}) \begin{pmatrix} 1 & x_{ij} \\ x_{ij} & x_{ij}^2 \end{pmatrix} \tag{4}$$

and

$$A = \sum_{i=1}^{m} \left\{ \sum_{j=1}^{n_i} w_{ij}^2 \mu_{ij} (1 - \mu_{ij}) \begin{pmatrix} 1 & x_{ij} \\ x_{ij} & x_{ij}^2 \end{pmatrix} + \sum_{j' \neq j} w_{ij} w_{ij'} \rho \sqrt{\mu_{ij}(1 - \mu_{ij}) \mu_{ij'}(1 - \mu_{ij'})} \begin{pmatrix} 1 & x_{ij'} \\ x_{ij} & x_{ij} x_{ij'} \end{pmatrix} \right\}, \tag{5}$$

where j and j' represent two arbitrarily chosen individuals from the i^{th} cluster and ρ refers to the within-cluster correlation with respect to the outcome, Y. (Note that we have made the assumption here that the intra-class correlation (ρ) is constant for all subjects, and not dependent on the value of covariates.) In certain cases, the expression for the $Var(\hat{\beta})$ simplifies. For example, suppose that the covariate of interest, X, is binary (e.g., presence/absence of exposure) and is cluster specific so that x_{ij} is the same for all members of the same cluster. Then, A and B simplify and in large samples will approximate the following:

$$B = \sum_{i=1}^{m} n_i E(w_i) \mu_i (1 - \mu_i) \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix}, \tag{6}$$

where $E(w_i)$ refers to the average of the weights for the i^{th} cluster, and

$$A = \sum_{i=1}^{m} n_i \mu_i (1 - \mu_i) \begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix} \left\{ E(w_{ij}^2 + \rho (n_i - 1) E(w_{ij} w_{ij'}) \right\}, \tag{7}$$

with $E(.)$ again referring to an average over the cluster. A few more special case considerations are helpful. First, consider the case where there is no within-cluster correlation (ρ=0) and also assume that the weights are independent of cluster membership and exposure, hence can be pulled out of the sums. It follows in this special case that

$$\mathrm{Var}(\hat{\beta}) = \frac{E(w^2)}{\left(E(w)\right)^2}\left\{\sum_{i-1}^{m} n_i \mu_i (1-\mu_i)\begin{pmatrix} 1 & x_i \\ x_i & x_i^2 \end{pmatrix}\right\}^{-1}. \qquad (8)$$

Note that this expression corresponds to the standard variance estimate based on a logistic regression, multiplied by a factor that involves the weights. The multiplicative factor can be re-expressed as:

$$\frac{E(w^2)}{\left(E(w)\right)^2} = \frac{\mathrm{Var}(w) + \left(E(w)\right)^2}{\left(E(w)\right)^2} = 1 + CV^2, \qquad (9)$$

or 1 plus the squared coefficient of variation of the weights. When the weights are constant, this factor equals 1 and the standard logistic regression variance formula applies. When the weights vary, then this factor will always exceed 1; hence the variance of parameters estimated using weighted estimating equations will always exceed those based on a simple logistic regression. This is a well known result among sample survey statisticians, and the extra term is often referred to as a *design effect.* These design effects provide a very useful tool when it comes to study planning and design, since one can think in terms of the impact of various different weighting schemes on the estimated variances of parameters of interest, and adjust accordingly.

Now consider the slightly more complex setting where the intra-cluster correlation, ρ, is non-zero. Using a similar logic, it is relatively straightforward to show the *design effect* (or the factor that multiplies the usual logistic regression variance) is:

$$1 + \rho(n-1) + CV^2 + \rho(n-1)\,\mathrm{cov}(w_{ij} w_{ij'}), \qquad (10)$$

where n is the average cluster size and the covariance term refers to the covariance between weights within the same cluster. In general, we would expect this covariance term to be 0. In the special case where the weights are all equal (variance and covariance of the weights equal 0), the design effect reduces to (1+ρ(m-1)), which is the usual inflation factor for a variance based on cluster data (see Diggle et al., 2002).

When the covariate of interest, X, is allowed to vary within-cluster, all these calculations become considerably more complicated. To facilitate our discussion here, consider the case where exposure is binary and let p_1 denote the probability that an individual is exposed, and p_0=(1-p_1) the probability that an individual is not exposed. Also, for simplicity, we define μ_1 to denote the response probability for exposed individuals and μ_0 the response probability for an unexposed individual. Then, it is relatively straightforward to show that the derivative of the estimating equation (see Equation (4)) will, in large samples, be approximately

$$B = mn\left[p_1\mu_1(1-\mu_1)\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix} + p_0\mu_0(1-\mu_0)\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}\right]E(w) = mn\Delta_1 E(w)\,, \qquad (11)$$

where, as before, n is the average cluster size and Δ_1 refers to the term in square brackets. Similarly, the variance of the estimating Equation (5) will be approximately:

$$A = mn\left[\Delta_1 E(w^2) + \rho(n-1)\Delta_2 E(w_{ij}, w_{ij'})\right], \qquad (12)$$

where Δ_2 is more complicated and equal to the following:

$$\begin{aligned}\Delta_2 = {} & p_{11}\mu_1(1-\mu_1)\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix} + p_{00}\mu_0(1-\mu_0)\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix} p_{11}\mu_1(1-\mu_1)\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix} + \\ & p_{10}\sqrt{\mu_1(1-\mu_1)\mu_0(1-\mu_0)}\begin{pmatrix}1 & 0\\ 1 & 0\end{pmatrix} + p_{01}\sqrt{\mu_1(1-\mu_1)\mu_0(1-\mu_0)}\begin{pmatrix}1 & 1\\ 0 & 0\end{pmatrix}\end{aligned} \qquad (13)$$

where p_{11} is the probability that two members of a cluster are both exposed, p_{00} is the probability that they are both unexposed, and p_{01} and p_{10} refer to the probability that one is exposed and the other is not exposed. In this complicated setting, it is not as straightforward to specify a design effect. However, consideration of some special cases for Δ_2 is worthwhile and allows us to explore the impact of various correlation patterns on estimated variances. First, consider the special case where there is perfect within-cluster correlation with respect to exposure values, meaning that p_{01} and p_{10} are both zero, $p_{11}=p_1$ and $p_{00}=p_0$. In other words this is once again the cluster-specific covariate case, where all subjects in a cluster are either exposed, or all subjects in a cluster are unexposed. In this case, Δ_2 is identical to Δ_1, and the design effect is once again given by equation (10). Alternatively, for the case where there is no within-cluster correlation with respect to X, we have $p_{11}=(p_1)^2$, $p_{00}=(p_0)^2$ and $p_{10}=p_{00}=p_0p_1$.

To derive more generic formulas for p_{00}, p_{01}, p_{10}, and p_{11}, we assume that the X's follow a beta-binomial such that

$$\Pr(X_{ij}=1|\pi_i) = \pi_i \qquad (14)$$

where π_i is the cluster-specific probability of response, and the π_i's follow a beta distribution:

$$\pi_i \sim \text{beta}(\alpha_1, \alpha_2). \qquad (15)$$

The beta parameters α_1 and α_2 are chosen so as to produce the desired marginal probability of exposure (p_1) and the desired within cluster correlation in the X's, λ. Under these assumptions we have $p_{00} = (1-p_1)^2+p_1(1-p_1)\lambda$, $p_{11} = p_1^2+p_1(1-p_1)\lambda$, and $p_{01} = p_{10} = p_1(1-p_1)(1-\lambda)$.

Using the above formulas, it is easy to use a computer package such as R or Splus to compute the variance of the estimated parameters under various assumptions on the degree of clustering and weighting. In particular, putting equations (11) and (12) together as in equation

(8) we get the approximate variance of the estimated parameter under the clustered and weighted design as:

$$V_{wc} = \frac{\left[\Delta_1^{-1} E(w^2) + \rho(n-1)E(w_{ij}, w_{ij'})\Delta_1^{-1}\Delta_2\Delta_1^{-1}\right]}{\left[E(w)\right]^2 mn}. \quad (16)$$

In contrast, the variance under simple random sampling is:

$$V_s = \left[\Delta_1^{-1}\right] / \left(mn\left[E(w)\right]^2\right). \quad (17)$$

Note that in general, there is no simple multiplicative relationship here, as we saw in the setting of cluster-specific covariates. Indeed, the relationship between variance estimates under simple and complex sampling differs according to which component of the parameter vector is being examined. To examine the ratio of the variances for the coefficient β_1 (i.e., the parameter that estimates the relationship between the health outcome and the exposure) we simply pull off the (2,2) elements of these two variance expressions and take their ratio. Additionally, note that the above formula depends on the number of clusters (m), the number of individuals in each cluster (n), the probability of exposure (p_1), the within cluster correlation in exposure (λ), the probability of the disease for unexposed individuals (μ_0), the within cluster correlation in the health outcome (ρ), the strength of the relationship between outcome and exposure (OR), and the degree of unequal weighting in the design.

A.3 References

Diggle, P., Heagerty, P., Liang, K-Y., Zeger, S. (2002). Analysis of Longitudinal Data. Oxford University Press.

Liang, K.Y. and Zeger S.L. (1986). Longitudinal data analysis using generalized linear models. Biometrika. 73:13-22.

Zeger, S.L. and Liang, K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 42:121-130.